Biotic Prediction

Building the Computational Technology Infrastructure for Public Health and Environmental Forecasting

Concept of Operations

BP-CONOP-1.0

Task Agreement: GSFC-CT-1

July 15, 2002

Contents

List of Tables

List of Figures

1 Overview

1.1 Introduction

This project will develop the high-performance, computational technology infrastructure needed to analyze the past, present, and future geospatial distributions of living components of Earth environments. This involves moving a suite of key predictive, geostatistical biological models into a scalable, cost-effective cluster computing framework; collecting and integrating diverse Earth observational datasets for input into these models; and deploying this functionality as a Web-based service. The resulting infrastructure will be used in the ecological analysis and prediction of exotic species invasions. This new capability will be deployed at the USGS Midcontinent Ecological Science Center and extended to other scientific communities through the USGS National Biological Information Infrastructure program.

1.2 Referenced Documents

Table 1. Referenced Documents

Document Title	**Version**	**Date**
Software Engineering / Development Plan	1.0	2002-04-08
Software Requirements Document	1.0	2002-07-15

1.3 Document Overview

This document, the *Concept of Operations*, describes the *Invasive Species Forecasting System* (ISFS) from an operational point of view. It is not intended to imply a design or convey implementation requirements for the functionality described herein.

It provides a brief introduction to the functional architecture of the system through its intended external interfaces and high level description of the functional elements that comprise the system.

Finally it describes the system's operational scenarios through a series of "use cases."

2 ISFS Functional Architecture

2.1 System Description

The ISFS system will have users from government agencies, universities, industry, and the general public. To the users, the ISFS is deployed as a web browser based system that will present options for applying a series of models to available datasets yielding predictive result sets. The system can ingest data from different sources and in different formats and existing models can be either run, or new ones created. The system outputs maps and additional information depicting the applied model and the predicted species distributions.

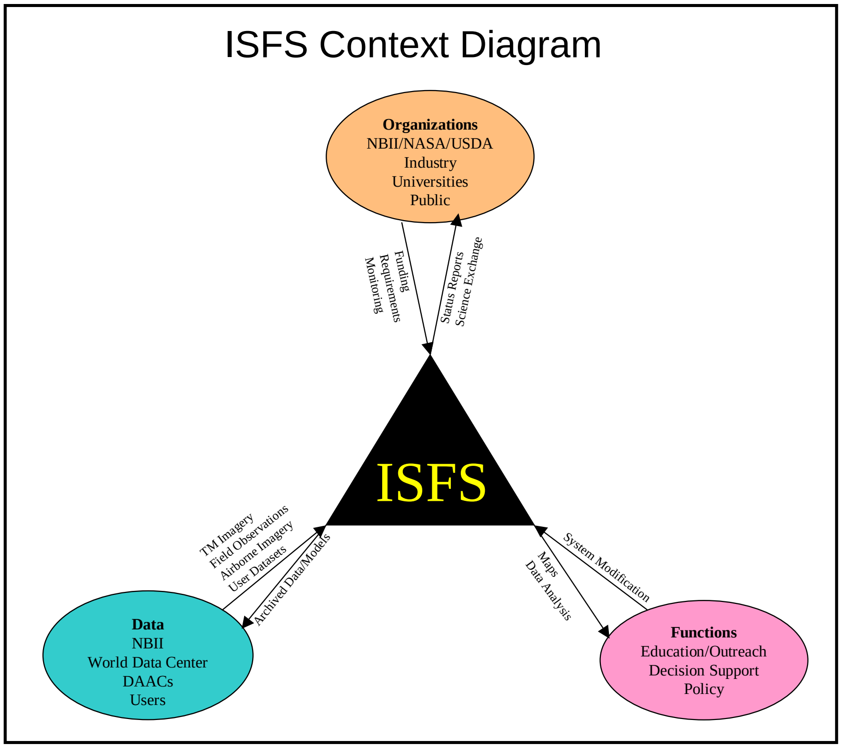


Figure 1. Context Diagram

2.2 External Interfaces

The primary means of interface to the ISFS will be a W3 standards compliant web browser. All GUI development efforts for external interfaces will be through web-based client architecture that uses HTTP as the primary means for interacting with the system. Supplmentary interfaces for ingest may include a secured FTP push/pull technique that will be scheduled through the GUI.

Standing orders for required data products will be accomplished through subscriptions to the varied data sources including the DAAC, NBII, USGS, and other data sources TBD. Different levels of access for public, model user, model builder, and developer will be established and documented in the SRD.

The user interface is implemented through an HTTP connection to browser-based GUI to select models and datasets, and to apply the selected model to the selected dataset. The GUI is also the vector for returning processed output to the end user.

2.3 Internal Architecture

The ISFS conceptually consists of 6 functional elements:

1. Ingest
2. Pre-processing
3. Archive
4. Modeling
5. Post-processing
6. User Interface

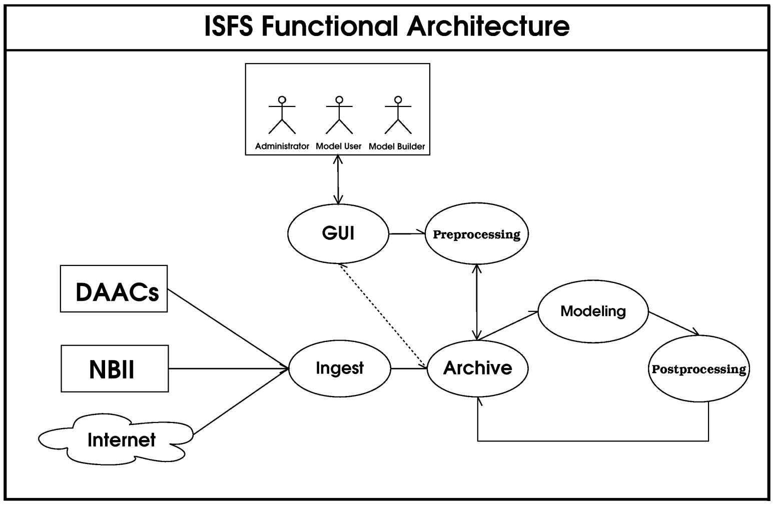


Figure 2. ISFS Functional Architecture

2.3.1 Ingest

The ingest subsystem will serve as the initial “entry point” for all data used in the system. The data fall into the three main categories: field point measurements, imagery, and ancillary layers. These categories will be further defined in the Requiremnts Analysis Document. Common to the three categories is that all data ingested into the system will be associated with some geographic location. There will be a validation step to verify the integrity of the data before ingest, and every effort will be made to ascertain that the data originated from an authoritative source.

Within the subsystem users will be able to upload field data in a tabular form using standard templates provided by the Invasive Species Forecasting System. These templates will ensure that all system required fields will be captured and will be in an accessible format (such as a spreadsheet, database, or simple ASCII list). Satellite data will be included primarily from external satellite data archives but also user-supplied satellite data or airborne imagery may be used. The primary source for ancillary layers will initially be USGS but here too the ingest system will allow user-supplied ancillary layers to be incorporated into the system.

Mechanisms for data acquisitions will be secured ftp-push for any user supplied data or automated secured ftp pull from the external archives. Users will be issued usernames and will be authenticated through passwords. User activity and preferences will be logged and archived in the system. A specific list of external archives and required data sets will be established and maintained as part of the ingest subsystem. The interfaces to these archives will be negotiated and a thorough understanding of source and target schema will be included in the interface agreement. The ingest subsystem will monitor the number and volume of data brought into the system, with an ability to break this down by location, user, and external archive.

For the purpose of establishing the baseline canonical example, we will assume that the datasets exist and are stored locally on our servers. Once the HPC technology has been applied and proven we can expand our ingest routines to include additional themes (ancillary layers) and interfaces with government and university databases. Formal procedures for ingest of these data will be documented as the data/source specific requirements become evident. As a general standard operating procedure we will capture metadata and QA type data that will describe each file to be ingested and write that into the file header or store it in a database record associated with the unique file identifier.

2.3.2 Pre-processing

The pre-processing subsystem merges the ingested datasets to create a data product that will be analyzed by the modeling subsystem. In the baseline scenario, the field data are merged with the Landsat and DEM (Digital Elevation Model) information at the same UTM x,y coordinates. The subsystem may perform resampling if the input data are not at the same resolution, and the Landsat data may be processed to higher level products, e.g. tassel cap coefficients, principal components, atmospherically corrected reflectance values, etc. The merged data product will be written to the archive in a common analysis format.

2.3.3 Archive

The archive subsystem will consist of a database that will store pointers to the archived files. The files that we maintain will be contained in a logically arranged directory structure and indexed with a unique file ID. For externally stored data, the archive system will store a file ID and pointer or URL that can be used to retrieve and stage the archived files for subsequent processing.

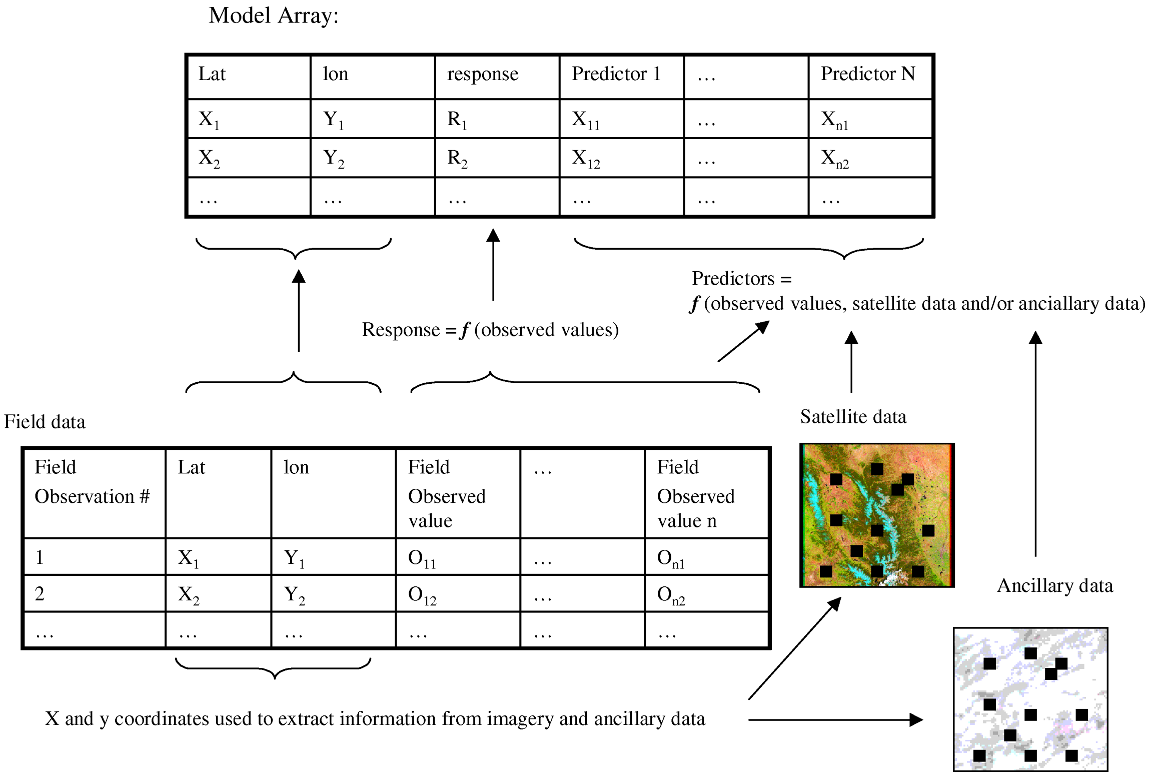
2.3.4 Modeling

The modeling will be empirical in nature and utilize statistical techniques. The general theme of the modeling will be to predict a certain species' migration through or invasion of habitat based on remote sensing imagery and ancillary data layers.

The modeling subsystem will consist of five components:

1. construction of response/dependent variable and predictor/independent variable data array
2. model selecting and fitting
3. model diagnostics
4. accepting model or adjustment/refinement (back to step 2)
5. model output

Each model will require an array containing the response, or dependent, variable (the “Y” variable) and a set of predictor, or independent, variables (the “X” variables). Constructing the data array needed for modeling will start with a set of geographic coordinates. These will likely come from the geographic coordinates associate with the tabular field data of interest. In addition to the coordinates, the Y variable will be extracted from the tabular field data, either directly or as a function of one or more elements in the field data. The X variables will come from any of the three data sources. X variables from the field data can also be directly extracted or be a function of one or more elements from the data. X variables will be extracted from the satellite and ancillary data by using the coordinate information from the field data to extract values from the imagery for the corresponding pixels. These X variables can come directly from the satellite or ancillary data or be a function of one or more satellite or ancillary variables. A diagram of the modeling array is shown in Figure 3.



Lat	lon	response	Predictor 1	…	Predictor N
X_1	Y_1	R_1	X_{11}	…	X_{n1}
X_2	Y_2	R_2	X_{12}	…	X_{n2}
…	…	…	…	…	…

Predictors =
f (observed values, satellite data and/or anciallary data)
Response = *f* (observed values)
Field data
Satellite data
Ancillary data

Field Observation #	Lat	lon	Field Observed value	…	Field Observed value n
1	X_1	Y_1	O_{11}	…	O_{n1}
2	X_2	Y_2	O_{12}	…	O_{n2}
…	…	…	…	…	…

X and y coordinates used to extract information from imagery and ancillary data

Figure 3. Schematic of Modeling Array

Once the array is constructed, the model selection will involve screening X variables to see which are most related to the Y variable. Methods to be considered are graphical exploratory analysis, stepwise regression, and combinatorial screening. The variables found to relate to the Y variable will be related through statistical models. Models to be considered fall into the generalized linear models framework using generalized least squares and regression tree models and others as applicable. Model diagnostics will include assessing the fit of the model and testing the assumptions implicit to the model. Of particular interest will be the spatial nature of the data and assumption of spatial independent or, alternatively, accounting for spatial dependence within the models. Results from the diagnostics will be used to either confirm the appropriateness of the model or influence adjustments or refinements to the model. Adjustments or refinements will require returning to model selection and fitting. Once an appropriate model is accepted, model output will include the model formula itself as well as metadata describing the user responsible for the model selection and the data used to drive the model.

2.3.5 Post-processing

The post-processing subsystem applies the results of the modeling subsystem to generate products that are then made available to the user. In the baseline scenario, the results of the OLS regression are applied to the input satellite and DEM data to provide a preliminary map of the total plants in the region. The kriged estimates of the residuals of this regression are then added to this map to produce an improved map. The subsystem will, at least, reproject the data as specified by the user and overlay with other data layers as requested. Finally, the data will be packaged with the appropriate metadata and assigned a unique data set identifier to be archived and made available to the user.

2.3.6 User Interface

The User Interface Subsystem provides a way of managing information and performing analyses by means of dynamically constructed activity-and-information spaces, or role-based views. The various role-based views are delivered through dynamically constructed, personalized Web pages.

A profile database maintains a profile of each user's status and preferences. Activities are implemented by a library of routines accessed through the controls on a web based forms interface. Each user is assigned a particular role, which enables them to perform certain actions but not others, and to view certain types of information but not others. Roles include the following:

- "Administrator" — provides complete access to the system,
- "Model Builder" — allows tailorability of the Ingest Subsystem and integration and use of new analytical routines to the Pre-Processing, Modeling, and Post-Processing Subsystems, and
- "Model User" — allows tailorability of the Ingest Subsystem and use of existing analytical routines.

3 Use Cases

We have used the “Use Case” Model [1] to describe the system and subsystems from the point of view of various actors interacting with the system.

3.1 Actors

The primary actors and their general goals are:

Administrator has complete access to the system, monitors performance, and coordinates and authorizes system design changes, etc.

Model Builder is able to build tailored datasets as well as new analytical routines.

Model User is able to build tailored datasets to be used by existing analytical routines.

[1] Cockburn, Alistair, *Writing Effective Use Cases*, Addison-Wesley Pub Co, 2000

3.2 Use Case 1: Run a pre-defined model

Primary Actor: Model User

Goal: User runs a pre-defined model.

Scope: ISFS

Level: User Goal

Preconditions:

1. User account exists on the system.
2. Input files are accessible by the system.
3. Desired analysis routine integrated into the system.

Trigger:

Success End Condition: Model User has run the model and generated appropriate results.

Failed End Condition: Model User unable to obtain results.

Main Success Scenario:

1. Model User logs in to the system and starts a session.
2. Model User selects input files from list of available options, selects desired analysis routine from available options, and hits “Run” button.
3. Model User receives the output predictive map and uncertainty map.
4. Model User receives the associated metadata describing output files, run parameters, and performance stats.
5. Model User can optionally save the run results with personal annotations in personal repository.

Extensions:

Open Issues:

3.3 Use Case 2: Create a new data file for analysis

Primary Actor: Model User

Goal: Model User creates new data file and documentation for subsequent analysis by the modeling subroutine.

Scope: ISFS / Pre-Processing Subsystem

Level: User Goal

Preconditions:

1. User account exists on the system.
2. Input files in a format understood by the system.

Trigger: A new file to be analyzed is available.

Success End Condition: User has created new data file and documentation for subsequent analysis.

Failed End Condition: User unable to create useable data file.

Main Success Scenario:

1. Model User logs in to system and opens a session and navigates to the Pre-Processing Subsystem and interface.
2. Model User merges ingested datasets to create a data product that can be analyzed by the modeling subsystem.
3. Model User generates metadata documentation for the new data file.
4. Model User can optionally save the new data file with personal annotations in personal repository.

Extensions:

Open Issues:

3.4 Use Case 3: Create a new analysis routine

Primary Actor: Model Builder

Goal: Model Builder creates a new analysis routine and documentation and integrates it into the modeling subsystem.

Scope: ISFS / Modeling subsystem

Level: User Goal

Preconditions:

1. User account exists on the system.
2. System software interfaces understood by the new analysis code.

Trigger: A new analysis routine is available.

Success End Condition: Model Builder has integrated new analysis routine into the modeling subsystem for subsequent use.

Failed End Condition: Model Builder unable to integrate new routine.

Main Success Scenario:

1. Model Builder logs in to the system, starts a session, and navigates to the Modeling Subsystem and interface.
2. Model Builder integrates new analysis routine into the modeling subsystem.
3. Model Builder integrates accompanying documentation for the new routine.
4. Model Builder can optionally test new analysis routine using existing data and output routines from the Pre-Processing and Post-Processing subsystems.

Extensions:

Open Issues:

3.5 Use Case 4: Create a new output product

Primary Actor: Model Builder

Goal: Model Builder creates a new type of output product and documentation and integrates it into the post-processing subsystem.

Scope: ISFS / Modeling subsystem

Level: User Goal

Preconditions:

1. User account exists on the system.
2. System software interfaces understood by the new output code.

Trigger: A new output routine is available.

Success End Condition: User has integrated new output routine into the post-processing subsystem for subsequent use.

Failed End Condition: Model Builder unable to integrate new routine.

Main Success Scenario:

1. Model Builder logs in to the system, starts a session, and navigates to the Post-Processing Subsystem and interface.
2. Model Builder integrates new output routine into the post-processing subsystem.
3. Model Builder integrates accompanying documentation for the new routine.
4. Model Builder can optionally test new output product using existing data and analysis routines from the Pre-Processing and Modeling subsystems.

Extensions:

Open Issues:

A Glossary

BP Biotic Prediction project

CT Computational Technologies project

CONOP Concept of Operations

COTS Commercial Off The Shelf

CSU Colorado State University

DAAC Distributed Active Archive Center

ESTO Earth Science Technology Office

FTP File Transfer Protocol

GSFC Goddard Space Flight Center

GUI Graphical User Interface

HTTP Hyper-Text Transport Protocol

ISFS Invasive Species Forecasting System

NBII National Biological Information Infrastructure

NREL Natural Resources Ecology Laboratory

SEP Software Engineering / Development Plan

USGS United States Geological Survey